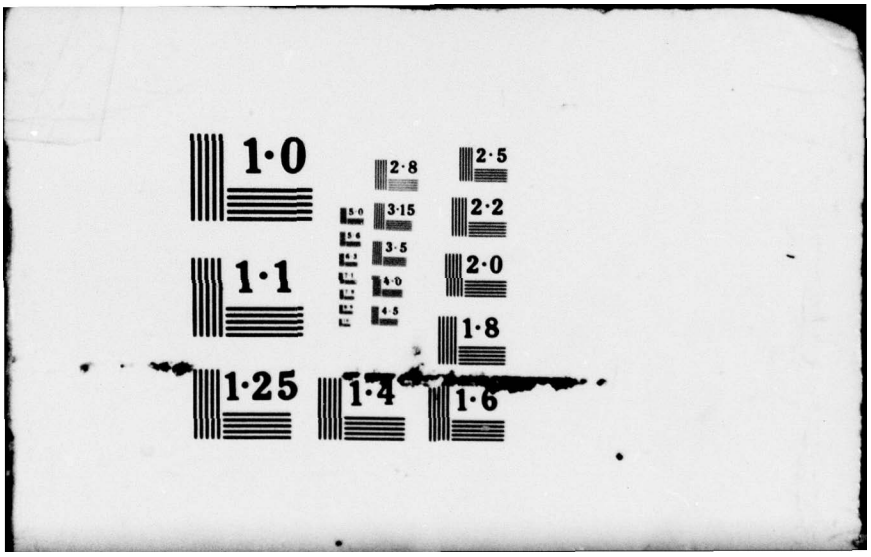
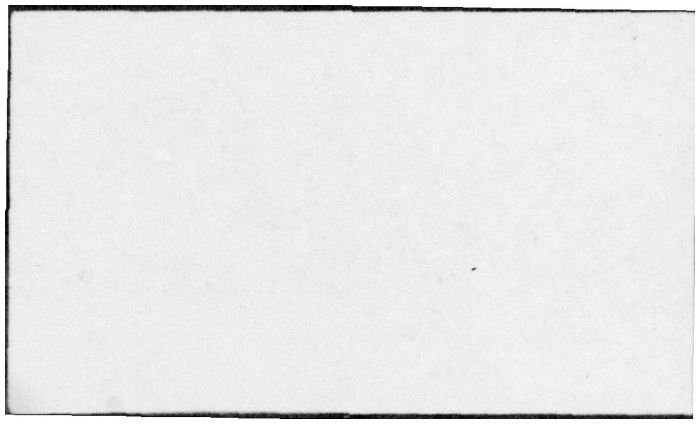
AD-A072 682 DESMATICS INC STATE COLLEGE PA F/G 12/1
STATISTICAL PROCEDURES FOR EXTRACTING OPTIMAL PREDICTOR VARIABL--ETC(U)
AUG 79 D E SMITH, J J PETERSON N00014-79-C-0128
UNCLASSIFIED TR-112-2 NL

1 OF 1
AD
A072682

END
DATE
FILMED
9-79
DDC


1·0
1·1
1·25
2·8
3·15
3·5
4·0
4·5
5·0
1·4
2·5
2·2
2·0
1·8
1·6



DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

STATISTICAL PROCEDURES FOR EXTRACTING OPTIMAL PREDICTOR VARIABLES FOR USE IN AN IMPACT ACCELERATION INJURY PREDICTION MODEL

by

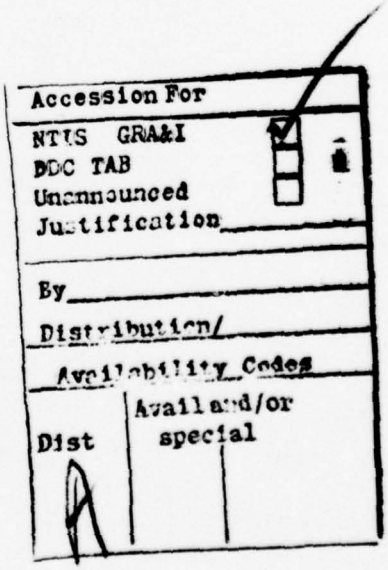
Dennis E. Smith
and
John J. Peterson

TECHNICAL REPORT NO. 112-2

August 1979

This study was supported by the Office of Naval Research
under Contract No. N00014-79-C-0128, Task No. NR 207-037

TABLE OF CONTENTS



I. INTRODUCTION

Previous Desmatics technical reports [2, 3, 5] investigated the use of a logistic function in the development of impact acceleration injury prediction models based on empirical data. The logistic models are of the form

$$P(\underline{x}) = \{1 + \exp[-(\beta_0 + \sum_{1}^{k} \beta_i x_i)]\}^{-1}$$

where

$\underline{x} = (x_1, \ldots, x_k)$ denotes the set of independent variables considered,

$(\beta_0, \beta_1, \ldots, \beta_k)$ denotes a set of unknown parameter values,

and $P(\underline{x})$ denotes the true probability of injury corresponding to $\underline{x}$.

Another report [4] described construction of "injury" (fatality) prediction models from actual $-G_x$ accelerator runs using subhuman primates (Rhesus monkeys) with restrained torso and unrestrained head. The data was obtained by the Naval Aerospace Medical Research Laboratory (NAMRL) Detachment as part of its research effort on impact acceleration injury prevention. Two prediction models were constructed from the data, each based on a different set of independent variables.

The first model was formed using three variables extracted from head dynamic response time trace data:

(1) peak head angular acceleration (resultant) measured in radians/sec^2,

(2) peak head linear acceleration (resultant) measured in meters/sec^2,

and (3) peak head angular velocity (resultant) measured in radians/sec.

The second model was based on two variables describing sled acceleration:

(1) peak sled acceleration measured in G's

and (2) rate of sled acceleration onset measured in G/sec.

Because of differing initial head positions of the experimental subjects, it was postulated a priori that the sled acceleration profile would yield less sensitive independent variables than head dynamic responses would. Using a common data base, two different models were constructed, one based on sled profile variables and the other based on head dynamic response variables. Although both models fit reasonably well, the model based on sled profile variables resulted in a much better fit [4].

It is intuitive that a model based on head dynamic response should provide predictions which are at least as good as those from a model based on sled profile. Thus, an explanation is required. A reason for the anomalous results could stem from inadequate extraction of information from head dynamic response time trace data.

To determine if, in fact, this is the case, care should be taken to insure that any set of variables describing head dynamic response comprises the best possible set of injury predictors. For the particular restraint configuration used, the data set used in model construction has shown the sled profile variables to be almost perfect predictors of injury likelihood. Thus, it is reasonable to extract, from the head dynamic response data, injury predictors that are highly correlated with the sled profile variables.

The extraction of such predictors from the head dynamic response time trace data can be achieved by the method of principal components. If there are several different kinds of time traces, it is desirable to condense the resulting predictor variables. The statistical method of canonical correlation analysis can be used to form an optimal condensation of these predictor variables with respect to the sled profile variables. The predictor variables

will be condensed in the form of two linear combinations canonically correlated with the sled profile variables. The statistical structure of principal components analysis and canonical correlation analysis is described in the following section.

II. STATISTICAL FORMULATION

Principal components are linear combinations of random variables which have special properties in terms of the variance. The first principal component is the normalized linear combination[1] of variables with maximum variance. The second principal component is the normalized linear combination of variables that have maximum variance among all linear combinations uncorrelated with the first principal component. The third principal component is the linear combination that has maximum variance among all linear combinations uncorrelated with the first principal and second principal components, and so forth. If there are n variables, it is possible to find n principal components, with each succeeding principal component having variance smaller than its predecessor. Usually the first few principal components will account for most of the variability in the data. Thus, these linear combinations usually contain most of the information in the data.

For each kind of dynamic response time trace, there is a corresponding set of principal components that contain most of the information in the data and, as such, define a set of potential predictors for injury likelihood. This set of predictors can be condensed by means of canonical correlation into two predictors in a way that describes the interrelationship between the sets of principal components and the sled profile variables.

Canonical correlation analysis is a statistical methodology used to express the interrelationships between two sets of variables. In this technical report, concern centers on the set of principal components

---

[1] The sum of squares of the coefficients equals one.

derived from the head dynamic response time trace data and the sled profile variables. The first canonical correlates derived from canonical correlation analysis are the linear combinations of variables in each set that have maximum correlation. The second canonical correlates are the linear combinations of variables in each set that have maximum correlation among those linear combinations uncorrelated with the first linear combinations. The number of variables in the smaller of the two sets is the maximum number of canonical correlates that exist.

Since the sled profile variable set contains only two variables, peak sled acceleration and rate of acceleration onset, there can only be two canonical correlate pairs. The parts of the canonical correlate pairs that are the linear combinations of the principal component set are the final predictors for injury likelihood. Because of the canonical correlation structure established between the principal components (which contain most of the head dynamic response information) and the sled profile variables (which are excellent predictors of injury likelihood), these final predictors should be good predictors of injury likelihood.

A. DATA PREPROCESSING

It is assumed that each head dynamic response time trace considered will, for each subject, be sampled at each of n equally-spaced time points. The observation sampled at time t for time trace i and subject j will be denoted by z_{ijt}. Figure 1 provides a diagrammatic view of the situation. In the analysis, the observations $z_{ij1},\ldots,z_{ijT}$ will be considered as comprising a T-dimensional vector $\underline{z}_{ij}$.

Before an attempt is made to apply the principal components/canonical correlation procedure, care must be taken to guard against unsatisfactory results because of the lack of preprocessing. Regardless of which head dynamic responses are selected for examination, peaks within corresponding time traces should be aligned, since it is reasonable to assume that peaks may be major contributors to injury. For example, if peak z linear acceleration were highly correlated with injury, its effect might not be noted if peaks were not aligned. In such a situation, the effect would be damped because of the peak occurring at different locations in the $\underline{z}_{ij}$ vectors.

Once peak alignment has been carried out, composite observational vectors as shown in Figure 2 can be formed by linking all head response vectors with the two corresponding sled profile variables. The values of these sled variables for subject j will be denoted by s_{1j} (peak sled acceleration) and s_{2j} (rate of sled acceleration onset).

B. DATA ANALYSIS

For each of the I dynamic response time traces, the corresponding data should be used to find the principal components. For every type of time

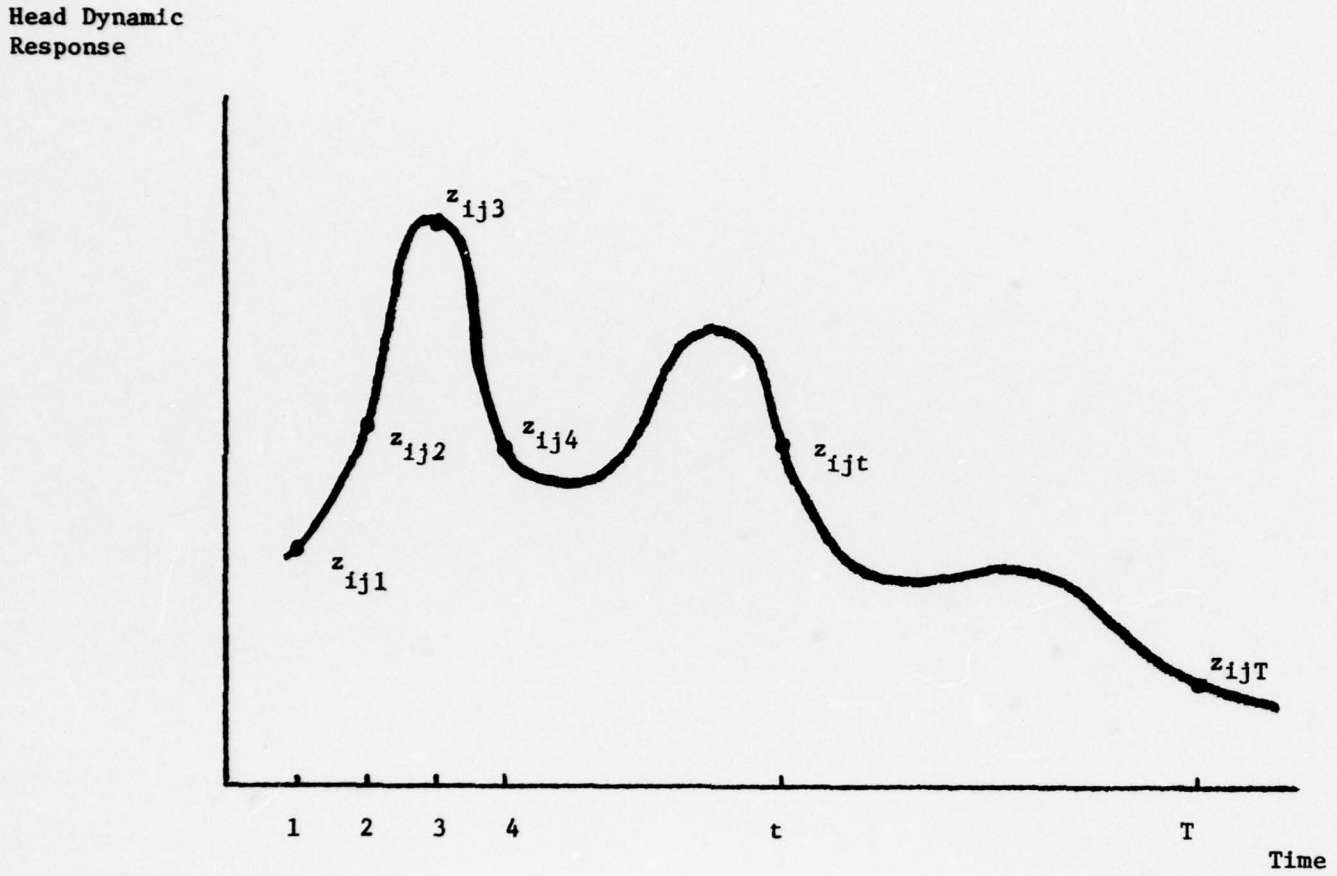


Figure 1: A Diagrammatic View of the Situation for Time Trace i and Subject j

Subject:	1	2	...	...	J
	$\begin{bmatrix} \underline{z}_{11} \\ \underline{z}_{21} \\ \vdots \\ \underline{z}_{I1} \\ s_{11} \\ s_{21} \end{bmatrix}$	$\begin{bmatrix} \underline{z}_{12} \\ \underline{z}_{22} \\ \vdots \\ \underline{z}_{I2} \\ s_{12} \\ s_{22} \end{bmatrix}$			$\begin{bmatrix} \underline{z}_{1J} \\ \underline{z}_{2J} \\ \vdots \\ \underline{z}_{IJ} \\ s_{1J} \\ s_{2J} \end{bmatrix}$

$\underline{z}_{ij}$ denotes the T-dimensional vector for time trace i and subject j

s_{1j} denotes peak sled acceleration for subject j

s_{2j} denotes rate of sled acceleration onset for subject j

Figure 2: Composite Observational Vectors

trace, enough principal components should be computed to account for most of the variability of the data in that time trace. Hopefully, the first few principal components should be enough to account for most of the variability of a particular time trace data set.

For each subject, the corresponding set of principal components scores of that subject's head dynamic response time traces can be concatenated along with the corresponding sled profile variables to form the observational vectors in the reduced data set as shown in Figure 3. There, for notational simplicity, M principal components are shown for each kind of time trace. If the number of subjects (J) is greater than the dimension (IM + 2) of these reduced observational vectors, than a canonical correlation analysis can be performed on this condensed data set.

Assuming a sufficient number of subjects, a canonical correlation analysis may be performed on the reduced observational vectors to obtain two linear combinations of the principal components that are correlated with the two sled profile variable linear combinations. The first canonical correlates are the linear combinations of the principal component set and the sled profile variable set that have maximum correlation with each other. The second canonical correlates are the linear combinations of the principal component set and the sled profile variable set that have maximum correlation among all those linear combinations uncorrelated with the first canonical correlates. The resulting two linear combinations of principal components, taken from the two canonical correlate pairs, are the injury likelihood predictors.

Subject:	1	2	...	...	J
	w_{111}	w_{121}			w_{1J1}
	.	.			.
	.	.			.
	.	.			.
	w_{11M}	w_{12M}			w_{1JM}
	w_{211}	w_{221}			w_{2J1}
	.	.			.
	.	.			.
	.	.			.
	w_{21M}	w_{22M}			w_{2JM}
	.	.			.
	.	.			.
	.	.			.
	.	.			.
	.	.			.
	w_{I11}	w_{I21}			w_{IJ1}
	.	.			.
	.	.			.
	.	.			.
	w_{I1M}	w_{I2M}			w_{IJM}
	s_{11}	s_{12}			s_{1J}
	s_{21}	s_{22}			s_{2J}

Figure 3: Observational Vectors in Reduced Data Set

III. COMPUTATIONAL PROCEDURE

The principal components and the canonical correlation analyses can be conducted with the aid of one of the several sophisticated computer packages currently available. In this section, the statistical package BMDP [1] is used for purposes of illustration. Within BMDP, the program BMDP4R can be used to perform the principal components analysis and the program BMDP6M can be used to perform the canonical correlation analysis. BMDP4R should be used on each time trace data set to obtain an output of the corresponding principal components. A subset of these outputs can then be properly arranged and input to the BMDP6M program to compute the canonical correlates.

The BMDP4R program computes the principal components and regresses them on a user specified dependent variable. The regression of the principal components on a dependent variable is not needed for the analysis described in this technical report. However, since the BMDP4R program must have a user specified dependent variable, it would be best to use "peak sled acceleration" in order to discern how the principal components are correlated with that sled profile variable alone. The input to each BMDP4R program for the $i^{\underline{th}}$ time trace consists of the vectors $\underline{z}_{ij}$ augmented with the peak sled acceleration s_{1j} for each of the J subjects. This input is pictorially shown in Figure 4.

Because the data within each dynamic response time trace is measured in the same units, no standardization of the time trace data is required. Therefore, the NO STANDARDIZE command should be used in the BMDP4R regression paragraph. Either of two criteria can be used to determine the order of

Subject	1	2	...	...	J
	$\begin{bmatrix} \underline{z}_{i1} \\ s_{11} \end{bmatrix}$	$\begin{bmatrix} \underline{z}_{i2} \\ s_{12} \end{bmatrix}$			$\begin{bmatrix} \underline{z}_{iJ} \\ s_{1J} \end{bmatrix}$

Figure 4: Input to the BMDP4R Program for the $i^{\underline{th}}$ Time Trace

entry of the principal components into the regression analysis. Specifying CORRELATION in the regression paragraph causes the principal components to be entered in the order of magnitude of the absolute value of their correlations with the dependent variable, the largest entered first. Specifying EIGENVALUE causes the principal components to be entered in the order of magnitude of the variance of the principal components. Regression on the principal components is not an important aspect of this analysis, but at this point interest centers on the magnitude of the variance of the principal components and specifying EIGENVALUE will cause the principal component coefficients to be conveniently output in the order of magnitude of the variance of the principal components.

The principal component coefficients will be denoted by c_{imt}. These are estimated parameter values. Specifying SCORE in the BMDP4R print paragraph causes the output of the principal component scores, i.e., the linear combinations of time trace points for each subject. The principal component scores will be denoted by w_{ijm}. If the first few principal components account for most of the variability of a given time trace data set, only the principal component scores corresponding to these first few principal components will be needed to form the reduced data set as shown previously in Figure 3. All the principal component scores can be output to a BMDP file. Note that the BMDP4R program does not output both of the sled profile variables along with the principal component scores. A BMDP4R program must be run for each type of head dynamic response time trace, and for each BMDP4R output, only the principal component scores corresponding to the first few principal components need to be used to form the data set shown in Figure 3. The data manipulation required to form the data set shown in that figure can be done manually or

by using BIMEDT, a Fortran transformation program of BMDP, in conjunction with output files created by the BMDP4R programs.

The BMDP6M program may be used to compute the canonical correlate pairs from the reduced data set shown in Figure 3. In the canonical paragraph the "FIRST" set of variables should be the principal component scores, for each of the time traces, taken from the reduced data set in that figure. The "SECOND" set of variables should be the sled profile variables. In the BMDP6M print paragraph the parameters COEF and CANV should be specified. COEF causes the output of the coefficients of the canonical correlates and CANV causes the output of the canonical correlate scores, i.e., the linear combinations of the principal components scores for each subject and the linear combinations of the sled profile variables for each subject. Let U_{1j} and U_{2j} represent the two linear combinations of principal component scores for the $j^{\underline{th}}$ subject. Let V_{1j} and V_{2j} represent the two linear combinations of the sled profile variables for the $j^{\underline{th}}$ subject. Then U_{1j} and U_{2j} are the candidates for the final predictors for injury likelihood for the $j^{\underline{th}}$ subject. (Figure 5 provides a summary of the overall procedure.)

$$(1)\quad w_{ijm} = \sum_{t=1}^{T} c_{imt} z_{ijt}$$

$$(2)\quad U_{1j} = \sum_{i=1}^{I} \sum_{m=1}^{M} a_{im} w_{ijm}$$

$$(3)\quad U_{2j} = \sum_{i=1}^{I} \sum_{m=1}^{M} b_{im} w_{ijm}$$

I is the number of time trace types.

M is the number of principal components used for each time trace.

a_{im}, b_{im} are coefficients estimated from the data set in Figure 3.

w_{ijm} is the $m^{\underline{th}}$ principal component score for the $j^{\underline{th}}$ subject, corresponding to the $i^{\underline{th}}$ time trace.

z_{ijt} is the $t^{\underline{th}}$ time trace point from the $i^{\underline{th}}$ time trace for subject j,

c_{imt} is a coefficient estimated from the z_{ijt} data in Figure 4.

U_{1j}, U_{2j} are the final predictors of injury likelihood for subject j.

Figure 5: A Summary of the Overall Procedure

IV. SUMMARY

Two statistical criteria are employed in defining the independent variables for the logistic prediction model. One is a high percentage of time trace variability associated with a principal component; the other is good correlation of the principal components with the sled profile variables. The logistic injury model is to be used, in part, to study how changes in head dynamic response affect changes in injury likelihood. The method of principal components is used to arrange and condense the head dynamic response time trace data into a form that accounts for the variability of the data in an optimal manner. The first few principal components of any one kind of time trace should account for most of the head dynamic response variability in that type of time trace.

Since the sled profile variables were found to be excellent predictors of injury likelihood, it is desirable to form statistics of the principal component data that are highly correlated with the sled profile variables. This can be achieved by the method of canonical correlation analysis. In short, the method of principal components is used to epitomize the head dynamic response time trace data and the method of canonical correlations is used to form statistics of the principal components that should predict injury likelihood well.

V. REFERENCES

[1] Dixon, W. J., ed., BMDP: Biomedical Computer Programs, University of California Press, Berkeley, CA, 1977.

[2] Peterson, J. J. and Smith, Dennis E., "Statistical Inference Procedures for a Logistic Impact Acceleration Injury Prediction Model," Technical Report No. 102-7, Desmatics, Inc., 1978.

[3] Smith, D. E., "Research on Construction of a Statistical Model for Predicting Impact Acceleration Injury," Technical Report No. 102-2, Desmatics, Inc., 1976.

[4] Smith, D. E., "An Examination of Statistical Impact Acceleration Injury Prediction Models Based on $-G_x$ Accelerator Data from Subhuman Primates," Technical Report No. 102-6, Desmatics, Inc., 1978.

[5] Smith, D. E. and Gardner, R. L., "A Study of Estimation Accuracy When Using a Logistic Model for Prediction of Impact Acceleration Injury," Technical Report No. 102-5, Desmatics, Inc., 1978.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 112-2

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): STATISTICAL PROCEDURES FOR EXTRACTING OPTIMAL PREDICTOR VARIABLES FOR USE IN AN IMPACT ACCELERATION INJURY PREDICTION MODEL

5. TYPE OF REPORT & PERIOD COVERED: Technical Report

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Dennis E. Smith and John J. Peterson

8. CONTRACT OR GRANT NUMBER(s): N00014-79-C-0128

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Desmatics, Inc.
P. O. Box 618
State College, PA 16801

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 207-037

11. CONTROLLING OFFICE NAME AND ADDRESS:
Biophysics Program (Code 444)
Office of Naval Research
Arlington, VA 22217

12. REPORT DATE: August 1979

13. NUMBER OF PAGES: 17

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): 27 pp

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

TR-112-2

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Impact Acceleration Injury Prediction
Predictor Variables
Principal Components
Canonical Correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An empirical impact acceleration injury prediction model can be based on an underlying logistic function using information extracted from dynamic response data to define independent (predictor) variables. This report describes statistical procedures for the extraction of optimal predictor variables. The application of the statistical techniques of principal components analysis and canonical correlation analysis is described. An outline of how the data analysis may be conducted with the BMDP statistical computer package is discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

391 156